A Multi-Modal Deep Learning Model for Drug Potency Prediction: Leveraging Features from Physics-Based Docking and Advanced Co-Folding Methods

Claire Suen
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, Ca, USA
claire.suen@merck.com

BoRam Lee
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, Ca, USA
boram.lee@merck.com

Yunsie Chung
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, Ca, USA
yunsie.chung@merck.com

Alan C. Cheng
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, Ca, USA
alan.cheng@merck.com

Merck & Co., Inc. and University of California Berkeley Collaboration*

Abstract

In drug discovery, the accurate prediction of a compound's potency is crucial for efficient design and optimization of small molecules as drugs. While machine learning and deep learning approaches can be useful, they generally require significant amounts of data that is not typically available in drug discovery programs in practice. We address this limitation by developing a multi-modal deep learning framework that enhances a graph neural network, Chemprop, by integrating explicit protein-ligand interaction features. We generated protein-ligand poses using both a physics-based docking method and two deep learning-based co-folding methods, Boltz-1 and Boltz-2. Our model demonstrates improved predictive accuracy for IC_{50} values for two diverse targets, CYP2D6 Inhibition and EGFR kinase. Additionally, our methods leveraging co-folding consistently outperforms the traditional docking-based approach. Feature selection analysis further revealed that π-stacking interactions were the most informative, appearing in the top-performing feature sets across all methods. In low-data regimes, the PLIP-informed models consistently outperformed established baselines. This work provides a scalable method to fuse complementary data modalities, offering both enhanced predictive performance and valuable mechanistic insights into drug-target interactions.

1 Introduction

The prediction of a compound's potency is crucial for the efficient design and optimization of small molecules as therapeutics. Given the cost and resource intensity of experimental potency

*Merck Co., Inc. (South San Francisco, CA): Matthew Adrian, Jeffery Zhou, Hyeyun Jung, David He; University of California, Berkeley (Berkeley, CA): Gean Hu, Kelly Hui, Aditi Jain, Qamil Mirza, Milena Novakovic, Joseph Park, Winston Qian, Aarav Shah, and Xina Wang

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science: The Reach and Limits of AI for Scientific Discovery.

measurements, more accurate computational affinity models offer a promising alternative with the potential to accelerate drug discovery[6]. Chemprop, a widely used graph neural network (GNN), demonstrates strong performance in prediction of molecular properties but operates primarily on 2D ligand characteristics [16]. This ignores the physical interactions at the atom and graph levels that govern binding affinity within the protein pocket [13, 17, 38].

However, machine learning approaches are data hungry, while prediction of compound potency is most useful early on in a drug discovery campaign when there are fewer compounds with potency measurements available, on the order of 500 or less. In this limit, GNNs such as Chemprop perform the same or worse than traditional machine learning approaches based on fingerprints, such as random forest or XGBoost [8]. This is because deep learning models require a large amount of data to learn a robust and generalizable representation of chemical space [42]. This presents a critical limitation and highlights the need for models that can achieve better predictive accuracy even with very little data.

Approaches for building improved models in lower data situations include pre-trained models such as MoLFormer or ChemBERTa [35, 36, 39], which has been found to generally perform similar to Chemprop for lower data situations [34]. Free-energy methods based on structure-based models show promising prediction accuracy but are compute intensive, typically taking on the order of hours or days for predictions, which is prohibitive for prioritizing large sets of idea molecules.

This suggests a fundamental challenge for deep learning models: without extensive pre-training, they essentially "re-learn" chemical intuition for each new task, necessitating large datasets to generalize effectively. This challenge highlights an opportunity to enhance the predictive performance of faster, ML-based models by integrating explicit 3D protein-ligand interaction information, in order to provide a richer, more task-specific context.

Furthermore, the increasing complexity of deep learning techniques often leads to "black-box" models. By incorporating explicit Protein-Ligand Interaction Profiler (PLIP) features, our approach links predictions to specific, tangible molecular interactions, thereby enhancing model interpretability. This provides not only improved performance but also mechanistic insights.

1.1 Main contributions

To address these limitations, our work enhances the predictive capabilities of Chemprop by integrating explicit protein-ligand interaction features (Figure A1). We first generate physically plausible binding poses for small molecules with EGFR and CYP2D6 Inhibition using molecular docking (GOLD) or a deep learning-based co-folding method (Boltz). We leverage PLIP to extract detailed atom-level features from these complexes, capturing specific physical interactions such as hydrogen bonds and hydrophobic interactions. The core of our approach lies in evaluating whether these PLIP-derived features can improve the accuracy of IC_{50} predictions for EGFR and CYP2D6 Inhibition within the Chemprop framework. In this work, we present a novel integration of protein-ligand interaction features within Chemprop. Our main contributions are summarized as follows:

Improved predictive accuracy: We demonstrate that incorporating features from protein-ligand interactions (π-stacking) improves the predictive accuracy for IC_{50} values. This highlights the role of protein binding context in molecular property prediction.

Identification of key features: Our model reveals that certain types of protein-ligand interactions are more informative than others for specific assays. For the targets we studied, we found that π-stacking interactions were particularly predictive and consistently appeared in the top-performing feature sets across all pose-generation methods.

Enhanced performance in low-data regimes: By incorporating explicit structural context, our π-stacking informed models consistently outperformed both the Chemprop and Random Forest baselines when trained on small subsets of the CYP2D6 Inhibition dataset, with statistical significance at the lowest training sizes.

1.2 Related Work

Molecular docking and co-folding methods Molecular docking (GOLD) is a physics-based method that predicts the optimal binding pose of a ligand by exploring a predefined search space [18]. Docking is limited by its reliance on a static protein structure, potentially missing key conformational

changes [4]. Therefore, co-folding methods, a class of deep learning models that predict the final protein-ligand complex structure by modeling the folding process in the presence of the ligand, present an opportunity for accurate 3D structure predictions [28]. Here, we focus on the Boltz models, a generative family of models that use a diffusion-based approach to predict protein-ligand poses [31, 44].

Protein-ligand interaction feature selection. Physics-based descriptors has been widely used due to their rule-based nature, providing interpretable and physically meaningful explanations for predictions [22]. These methods characterize binding pockets and engineer features from protein sequences, contacts or more detail interactions that can be extracted using tools like PLIP, OpenBabel and others [2, 13, 30, 45]. Advances in machine learning have introduced empirical scoring functions derived from, but not limited to, learned embeddings, embedding-based potential energy estimations, and interaction-aware network such as mixture density network [9, 15, 20, 23, 37]. These developments enable the use of diverse descriptors, showing promising improvement in the accuracy of binding affinity predictions. However, machine learning-derived features often require additional model training, which can increase computational time and resource demands.

Protein-ligand binding affinity prediction. Protein-ligand binding prediction methods can be broadly categorized into two types: foundation models and target-specific models [22, 31]. Foundation models aim to provide generalizable predictions across diverse targets by leveraging large datasets and broad applicability [33]. In contrast, target-specific models tailor their predictions to unique characteristics of a particular target, often achieving higher accuracy with less data and computational resources. Zero-shot and few-shot learning approaches have recently gained attention for their potential to improve generalization in low-data regimes, though there remains room for improvement [21]. A variety of architectures have been developed for protein-ligand binding affinity predictions [1, 10, 25]. Among these, GNNs have become popular in related tasks due to their ability to represent molecular structures as graphs, effectively capturing both chemical properties and spatial relationships [26, 41]. Typically, ligands are represented as graphs, while proteins are encoded either as sequences or graphs [29, 24]. Several studies have extended GNNs by integrating recurrent neural networks, graph isomorphism network, and transformer architectures [40, 46].

2 Methods

Our methodology encompasses four main stages: input preparation, protein-ligand structure prediction using molecular docking or co-folding, protein-ligand interaction profiling, and integration within a D-MPNN architecture. Dataset and input preparation method information can be found in Appendix A.1.

2.1 Protein-ligand structure prediction

We tested three approaches for structure prediction: molecular docking, co-folding with Boltz-1 or Boltz-2, and co-folding followed by energy minimization.

Molecular docking. Molecular docking simulations were performed using GOLD to predict the binding poses of the prepared ligands within the active sites of EGFR and CYP2D6 Inhibition [18]. For each ligand, GOLD generated 5 distinct docked poses. The pose with the most negative binding affinity score (representing the strongest predicted binding) was selected as the representative binding conformation for further analysis.

Protein-ligand conformation generation with Boltz. Separately, we employed a co-folding method (Boltz-1 and 2) to generate protein-ligand conformations for our study [31, 44]. The Boltz-generated structures were further refined using the Molecular Operating Environment (MOE, version 2024.06) [7]. The post-processing involved running the QuickPrep function with default settings. This procedure ensured the structural integrity of the generated poses by protonating the structure, adding missing hydrogens, correcting distant atoms, and performing energy minimization. Additional details on the Boltz structure predictions can be found in Appendix A.1.3.

2.2 Protein-ligand interaction profiling.

The protein-ligand complexes obtained by docking and co-folding were analyzed using Protein-Ligand Interaction Profiler (PLIP) to identify atom-level interactions [2]. PLIP generated various

interaction types, including hydrogen bonds, hydrophobic interactions and π-stacking. Two types of interaction features were extracted for each ligand.

Binary atom-level interaction vector (B). For each ligand, a binary vector is generated with a length corresponding to the number of atoms, where a value of 1 indicates the presence of an interaction and 0 indicates its absence.

Weighted atom-level interaction vector (C). Similar to the binary vector, this vector is generated with a length corresponding to the number of atoms, but each index is a continuous value weighted by the distance to the corresponding protein atom.

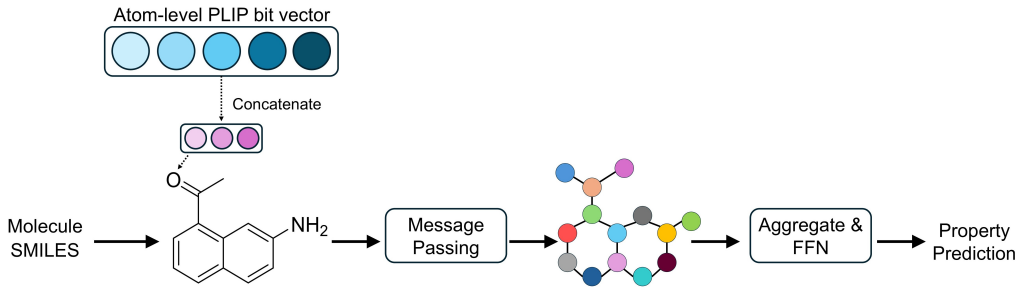


Figure 1: Integration of protein-ligand interaction features into the D-MPNN (Chemprop) architecture. The model takes two inputs: molecule SMILES string and a vector of specified protein-ligand interaction features. These features are concatenated with the atom-level features of the molecular graph prior to the message passing steps.

2.3 Integrate into D-MPNN architecture

The atom interaction features will be concatenated with the existing atomic features of each ligand atom. This concatenation will occur at the graph construction layer of Chemprop, prior to the message passing steps (Figure 1). We aim to enrich the atomic representations with critical protein-ligand contextual information at the foundational level, thereby enabling the model to learn more nuanced and context-aware relationships between molecular structure, protein-ligand interactions, and the resulting IC_{50} values. Additional details on methods are found in Appendix A.1.

3 Results

3.1 Feature selection

To identify the most informative protein-ligand interaction features, we performed a systematic feature selection process. For both the public CYP2D6 Inhibition and EGFR datasets, the training set was partitioned into training/validation subsets (random split) and a left out test set used for final performance statistics. We then trained a separate model for each feature extraction method (Docking, Boltz-1, Boltz-1 & MOE, Boltz-2, and Boltz-2 & MOE) on the partitioned training set. The models were evaluated on the validation set across all possible feature combinations. From these validation results, shown in Table 1, we selected the top five feature combinations for each method based on predictive performance. The selected top features showed that π-stacking (binary) was present in the top five combinations for all five methods. Hydrogen bonding (binary and continuous) and π-stacking (continuous) was present in the majority of the methods.

3.2 Performance of PLIP informed model

The baseline D-MPNN (Chemprop) and Boltz models used two folds with two ensembles each ($2x2$), where the final prediction was the average of four individual model predictions [12]. The standard error for the baseline was calculated from the individual metrics of these four models. Our PLIP-informed models, incorporating the selected features, were evaluated against this baseline. The Docking model predictions were obtained by training a model each of the selected RCSB structures

Table 1: Feature selection results for the public CYP2D6 inhibition dataset. Summary of the most informative protein-ligand interaction features identified through a systematic evaluation of five different pose-generation methods. A checkmark (✓) indicates that a given feature or combination of features was present in at least one of the top five performing models for that method on the validation set. **Bold features** appear in the majority (at least 3 out of 5) of methods.

CYP2D6 Inhibition	Docking	Boltz-1	Boltz-1 (MOE)	Boltz-2	Boltz-2 (MOE)
Hbond (B)			✓	✓	✓
Hbond (C)	✓	✓	✓		✓
Hydrophobic (B)	✓				✓
Hydrophobic (C)			✓	✓	
π-Stacking (B)	✓	✓	✓	✓	✓
π-Stacking (C)	✓	✓		✓	✓
π-Stacking (C), Hydrophobic (B)	✓			✓	
π-Stacking (B), Hydrophobic (B)		✓			
π-Stacking (B), Hydrophobic (C)		✓			
π-Stacking (B), Hbond (B)			✓		

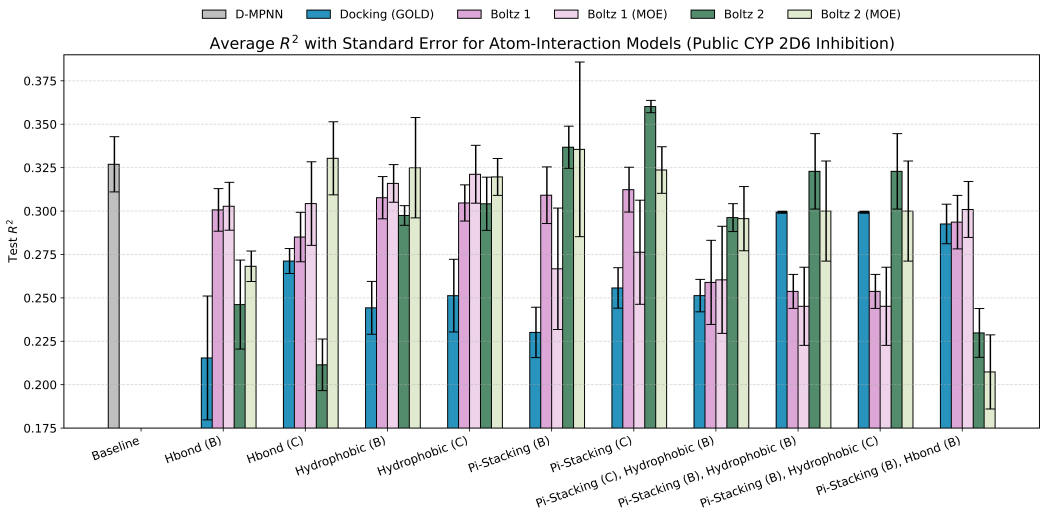
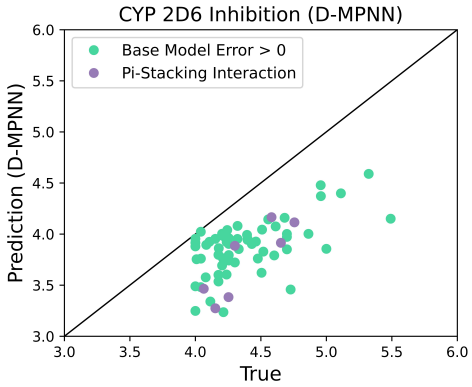


Figure 2: Performance of PLIP-Informed Models for Public CYP2D6 Inhibition. The bar chart compares the average R^2 with standard error for the baseline D-MPNN (Chemprop) model against selected PLIP-informed models.

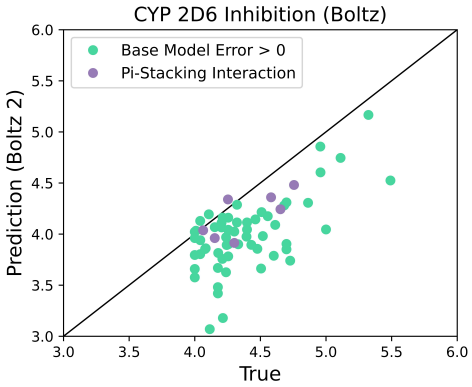
(two structures for CYP2D6 inhibition and one structure for EGFR) and obtaining the average prediction. The standard error was obtained by the individual metrics for these two models.

Public CYP2D6 Inhibition. Evaluating our models on the public CYP2D6 inhibition dataset (test n = 189), we found that the Boltz methods consistently outperformed the traditional docking-based approach, as shown in Figure 2. The Boltz-2 model that incorporates continuous π-stacking features outperformed all other models, including the baseline D-MPNN. Additionally, for models involving hydrophobic and hydrogen bonding interactions, those that were post-processed with MOE overall demonstrated performance improvements compared to their Boltz-alone counterparts. While the Boltz-2 method identified the least number of molecules with π-stacking interactions compared to other methods ($n = 155$ versus $\sim n = 500$), the KDE plots (Figure A5, Figure A6) showed that all methods followed a similar distribution for these interactions. This distribution roughly follows the shape of the overall CYP2D6 Inhibition histogram (Figure A10), indicating that π-stacking is an informative feature present across all potency levels.

Public EGFR. Similar results were found when evaluating our methods on the public EGFR dataset (test n=804). As shown in Figure A4, the Boltz-1 model with continuous π-stacking features



(a) DMPNN Parity Plot



(b) Boltz-2 (Continuous π-Stacking) Parity Plot

Figure 3: Comparison of predicted versus true potency for CYP2D6 Inhibition. The figure provides a side-by-side comparison of parity plots for the baseline D-MPNN model (a) and the Boltz-2 (continuous π-stacking informed) model (b) on the public CYP2D6 Inhibition dataset. A molecule's position on the parity line (True = Predicted) indicates perfect predictive accuracy. The figure highlights that the D-MPNN baseline model consistently underpredicted the potency of less potent compounds (those with a $log(IC_{50}) > 4$). In contrast, the Boltz-2 model corrects some of these underpredictions.

outperformed the D-MPNN baseline model. Overall, the Boltz-1 method consistently outperformed the traditional docking-based approach and the Boltz-2 method on most interaction models.

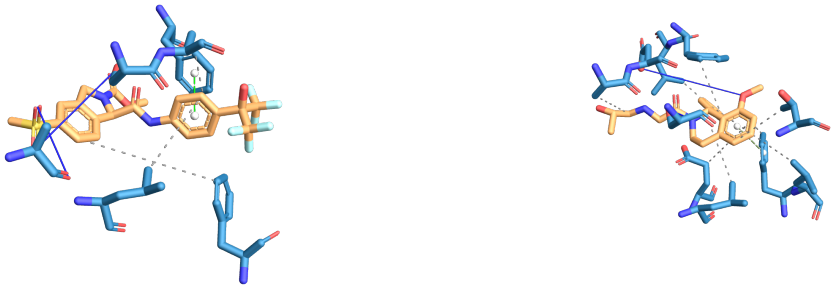
3.3 Analysis of π-stacking interactions

Public CYP2D6 Inhibition. We are interested in the specific role of π-stacking that is able to improve the CYP2D6 Inhibition predictive ability. We generated parity plots (Figure 3) for both the baseline D-MPNN model and the Boltz-2 (continuous π-stacking) model. Our analysis plots show the molecules where the D-MPNN baseline model performed poorly, specifically those with a true value greater than four (on the log-transformed scale) and an error greater than zero, indicating an under prediction of the true activity. When these same molecules were plotted on the parity plot for the Boltz-2 model, the predictions were generally closer to the parity line, demonstrating a decrease in prediction error for this subset of less potent compounds. The second group of molecules that were highlighted were molecules in the test set that were identified to have π-stacking interactions. The predictions for these molecules in the Boltz-2 model were consistently closer to the parity line when compared to their prediction value on the baseline model plot.

In Figure 4 we examined two compounds (Compound 1791 and 1834) where their true value is greater than four and identified to have π-stacking. From the generated PLIP interaction they are both interacting with the same residue: Phe120. Cross-checking across the other methods (Boltz-1, Boltz-1 MOE, and Boltz-2 MOE), Boltz-1 MOE also identified π-stacking for Compound 1791 and Boltz-2 MOE identified π-stacking for Compound 1834. However, the Boltz-1 MOE prediction for Compound 1791 and Boltz-2 MOE prediction for Compound 1834 were not improved. Boltz-1 MOE and 2 MOE both identified the same residue (Phe120) for the π-stacking interaction.

3.4 Performance in low-data regimes

To evaluate the effectiveness of our PLIP-informed model in data-scarce scenarios, a challenge in early stage lead design, we trained models on small subsets of the public CYP2D6 Inhibition dataset ($n = 250, 500, 750$) and tested their performance on the full test set. We chose to use the features that showed promise from Figure 2: π-stacking (binary and continuous). As shown in Table 2, we conducted a series of independent samples t-tests for pairwise comparisons between the π-stacking (binary and continuous) against the three baseline models (Chemprop, Random Forest, and MoLFormer). Given the multiple comparisons performed, there was an increased risk of a Type 1 error (false positive). To control the Familywise Error Rate (FWER) and ensure the reliability of



(a) Compound 1791 (b) Compound 1834

Figure 4: Correction of under predicted potency in π-stacking (green dashed line) identified compounds. A comparison of the predicted versus true IC_{50} values for two representative molecules from the test set. Compound 1791 (True = 4.14) was under predicted by the baseline D-MPNN model (Predicted = 3.27), but the Boltz-2 (continuous π-stacking informed) model corrected the prediction to 3.96. Similary, for Compound 1834 (True = 4.25), the Boltz-2 model (Predicted = 4.34) demonstrared a more accurate prediction than the baseline (Predicted = 3.38).

Table 2: Performance of PLIP-informed models in low-data regimes compared to benchmark models. π-Stacking (B and C) features are obtained from the Boltz-2 model. Average test R^2 values and standard errors are shown for models trained on compound subsets ('cmpd') of the public CYP2D6 Inhibition dataset. Top performing model is **bolded** for each training subset.

Model	250 cmpds	500 cmpds	750 cmpds
Chemprop	-0.04 ± 0.011	0.16 ± 0.033	0.24 ± 0.009
Random Forest	-0.03 ± 0.016	0.11 ± 0.019	0.18 ± 0.026
MoLFormer	-0.01 ± 0.011	0.03 ± 0.012	0.23 ± 0.019
π-Stacking (B)	$\mathbf{0.08 \pm 0.017}$	$\mathbf{0.22 \pm 0.011}$	0.24 ± 0.031
π-Stacking (C)	$\mathbf{0.08 \pm 0.013}$	$\mathbf{0.21 \pm 0.031}$	0.22 ± 0.039

our findings, we applied a Bonferroni correction to the significance threshold, which required the individual p-values to be less than $\frac{\alpha}{m}$ (where $\alpha = 0.05$ and m is the number of comparisons). Since we performed $m = 6 * 10 = 60$ pairwise comparisons, the required individual p-values had to be less than 0.00083. The PLIP-informed models consistently outperformed both the Chemprop, Random Forest, and MoLFormer baselines for $n = 250, 500$ and was equivalent in performance for $n = 750$. At the lowest training size ($n = 250$), the π-stacking models demonstrated a statistically significant improvement over the best-performing baseline, Random Forest. The π-stacking (continuous) model significantly outperform the Random Forest model (T-statistic: -6.81, p-value = 0.0005). Additionally the Boltz-2 binding affinity predictions performed significantly worse, $R^2 = -17.45$.

4 Discussion

Our modeling of the public CYP2D6 Inhibition and EGFR datasets using several benchmark and interaction-informed models lead to key findings around interaction feature importance, the low data regimes where addition of PLIP features notably improve predictive performance, and the role of co-folding methods in predictive performance.

4.1 Identification of π-stacking features

The consistent presence of π-stacking (binary and continuous) in the top-performing feature sets across all pose-generation methods underscores its importance as a key determinant of binding affinity. While other features like hydrogen bonding and hydrophobic interactions are also relevant, the increased performance of models leveraging π-stacking suggests a highly specific and geometrically constrained role for this interaction. Our parity plot analysis showed that for molecules where

the baseline model under predicted activity (i.e., less potent compounds), the Boltz-1 model with π-stacking features reduced the prediction error. This suggests that the model is learning to identify a specific structural characteristic, the presence of a π-stacking interaction, that correlates with a more favorable binding mode and correspondingly improved potency. KDE plots of π-stacking interaction frequency versus potency (Figures A5, A6) show consistency in distributions of π-stacking interactions across different pose-generation methods, suggesting that the underlying biological importance of π-stacking interactions is consistently captured.

For molecules with predicted π-stacking interactions, both Boltz-1 and Boltz-2 identify an increased frequency of π-stacking for more potent molecules in the single digit nM affinity range. Boltz-2 also flags fewer instances of π-stacking overall. These results suggest that for π-stacking interactions, Boltz-2 models, and Boltz-1 to a lesser degree, are better able to model critical π-stacking interactions while being more discriminate in predicting π-stacking interactions, and these features can lead to the observed improvement in predictive performance.

The addition of the π-stacking feature acts as an additional descriptor that compels the underlying model architecture to learn a more generalized relationship of the protein-ligand binding pocket. With an estimated binding energy of ~ 3 kcal/mol, π-stacking provides a substantial energetic contribution that can lead to a ~ 100-fold improvement in potency [11, 32]. This is significantly greater than the contributions of hydrogen bonds ($\sim$ 1 kcal/mol, $\sim$ 5x potency) and hydrophobic interactions (~ 0.25 kcal/mol, ~ 1.5x potency). The model's ability to predict this high-impact interaction appears to enrich its overall understanding of the binding environment. This allows it to generate more accurate predictions for molecules that lack a π-stacking interaction by learning a more context-aware representation of the protein's binding pocket. The greater relative strength of π-stacking interactions makes it an important interaction to capture for potency prediction, and this type of topological interaction is arguably more difficult to capture with a model operating purely in 2D space compared to models that account for 3D spatial relationships.

We note that the importance of π-stacking features may be target-specific. While our study demonstrates that π-stacking is a crucial and highly predictive feature for both CYP2D6 and EGFR inhibition, this may not be universally true for all protein targets. The observed significance is likely a result of the specific architecture of the binding pockets of these two targets, which contain key aromatic residues (Figure 4) that can engage in strong π-stacking interactions. For other targets, where the binding pocket is dominated by other features such as a network of hydrogen bonds or extensive hydrophobic interactions may be the primary determinants of binding affinity. This highlights the need for further investigation across a diverse range of protein families to fully understand the generalizability of our findings.

4.2 Low-data regimes

To understand the performance of the benchmark models and our PLIP-informed models in low data regimes commonly found in drug discovery settings, we took the CYP2D6 Inhibition dataset and created 250, 500, and 750 compound subsets (random select from train set) that represent typical data regimes for drug discovery programs. A typical small molecule discovery program generates on the order of 1000 to 2000 compounds in lead optimization, and potency models are generally more useful earlier on when there are fewer compounds synthesized and tested for on-target potency.

Comparing the Chemprop, random forest, and MoLFormer models with the PLIP-informed models suggests that the PLIP-informed models are particularly useful in lower data regimes, where they have significantly better performance. At 250 and 500 compound training set sizes, the PLIP-informed models clearly outperform the baselines. At a 750 compound training set size, the models perform about equivalently. Taken together, this suggests that the PLIP-informed models are more useful in low-data situations commonly found early on in drug discovery campaigns. We also see an improvement in predictive performance for MoLFormer at the 500 compound training set size, although the improvement is slight and not as large as with PLIP-informed models.

4.3 Co-folding versus docking

We initially used docking with the GOLD software package to generate binding pose models, and generally did not see predictive performance improvements. With the use of recent Boltz-1 and Boltz-2 co-folding models, we are starting to see improvements in PLIP-informed models (see Figure 2).

We also find that minimization of Boltz-1 or Boltz-2 models using a force-field (as implemented in MOE; see Methods) leads to greater variability in predictive performance.

We analyzed predicted protein-ligand complexes by comparing those from molecules similar to known PDB structures. Generally we found Boltz-2 showed best alignment to the PDB core substructure (Figure A2, A3).

5 Limitations & future directions

While our initial findings show promise, we need to test our approach on additional drug discovery targets. The improvements observed might be tied to the dominant role of π-stacking in these particular protein-ligand systems. We are actively looking at more targets to validate our methodology's broader applicability and reproducibility. Additionally, the overall performance in the most extreme low-data regime ($n = 250$) remains a challenge, with absolute R^2 values that are not yet ideal, though representing a statistically significant improvement over baselines. We believe that a major bottleneck lies in the quality of the generated protein-ligand poses. The methodology is dependent on the accuracy of the upstream co-folding and docking methods. If these methods fail to produce a correct binding pose, the PLIP-derived features will be based on inaccurate structural data, effectively introducing noise that can limit the model's predictive power. Future work will focus on integrating more advanced pose-generation techniques to provide a more reliable foundation for feature extraction and model performance.

Acknowledgments and Disclosure of Funding

The authors thank members of the Modeling and Informatics group at Merck & Co., Inc., for helpful discussions and the UC Berkeley CDSS Data Discovery Program for facilitating the collaboration. This work was partly funded by Merck & Co., Inc. (Rahway, NJ, USA).

References

[1] Gelany A. Abdelkader and Jeong-Dong Kim. Advances in protein-ligand binding affinity prediction via deep learning: A comprehensive study of datasets, data preprocessing techniques, and model architectures. *Current Drug Targets*, 25(15):1041–1065, 2024. ISSN 1389-4501/1873-5592. doi: 10.2174/0113894501330963240905083020.

[2] Mateo Adasme, Manuel Zurita-Gutiérrez, Carolina Fuentes-Martínez, María Fernanda Díaz-Oyarce, Felipe A. Sanhueza-Olivares, Javiera Gutiérrez-Jaramillo, Leonardo J. Rojas, Enrique Baez-Navarro, Javiera Peñaloza, Edwin Orozco-Guillén, María Macarena Gutiérrez-Abarca, Rodrigo Gutiérrez-Navarro, Karen J. Orellana, Cristóbal Martínez-Pérez, Francisco Bravo-Moraga, Víctor A. Díaz-Maldonado, Carlos Sotomayor-Pantoja, José M. Orozco, Nicolás Morales-Gómez, Diego Gutiérrez-González, Marcial Sáez-Sáez, Eduardo Paredes-Osses, and Carlos Pardo-Ortiz. PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Research*, 49(W1):W405–W410, 2021. doi: 10.1093/nar/gkab294.

[3] Matthew Adrian, Yunsie Chung, and Alan C. Cheng. Denoising drug discovery data for improved absorption, distribution, metabolism, excretion, and toxicity property prediction. *Journal of Chemical Information and Modeling*, 64(16):6324–6337, 2024. doi: 10.1021/acs.jcim.4c00639. PMID: 39108185.

[4] P. C. Agu, C. A. Afiukwa, O. U. Orji, et al., and Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Sci Rep*, 13:13398, 2023. doi: 10.1038/s41598-023-40160-2.

[5] Carolyn R. Butler, Kelly Ogilvie, Lidia Martinez-Alsina, Gonzalo Barreiro, Elizabeth M. Beck, Christopher E. Nolan, Kenneth Atchison, Elizabeth Benvenuti, Leigh Buzon, Stephanie Doran, Clarissa Gonzales, Christopher J. Helal, Xiaocheng Hou, Mei-Hsiu Hsu, Eric F. Johnson, Kevin Lapham, Laura Lanyon, Karen Parris, Brian T. O'Neill, Dale Riddell, Anthony Robshaw,

Felix Vajdos, and Michael A. Brodney. Aminomethyl-derived β-secretase (bace1) inhibitors: Engaging gly230 without an anilide functionality. *Journal of Medicinal Chemistry*, 60(1): 386–402, 2017. doi: 10.1021/acs.jmedchem.6b01451.

[6] Elena L. Caceres, Matthew Tudor, and Alan C. Cheng. Deep learning approaches in predicting admet properties. *Future Medicinal Chemistry*, 12(22):1995–1999, 2020. doi: 10.4155/fmc-2020-0259.

[7] Chemical Computing Group ULC. Molecular operating environment (MOE), 2024.0601, 2025. 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7.

[8] Jacky Chen, Yunsie Chung, Jonathan Tynan, Chen Cheng, Song Yang, and Alan C. Cheng. Data scaling and generalization insights for medicinal chemistry deep learning models. *Journal of Chemical Information and Modeling*, 2025.

[9] Shihong Chen, Haicheng Yi, Zhuhong You, Xuequn Shang, Yu-An Huang, Lei Wang, and Zhen Wang. Local–global structure-aware geometric equivariant graph representation learning for predicting protein–ligand binding affinity. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):15181–15193, 2025. doi: 10.1109/TNNLS.2025.3547300.

[10] Xiying Chen, Jinsha Huang, Tianqiao Shen, Houjin Zhang, Li Xu, Min Yang, Xiaoman Xie, Yunjun Yan, and Jinyong Yan. Deattentiondta: protein–ligand binding affinity prediction based on dynamic embedding and self-attention. *Bioinformatics*, 40(6):btae319, 06 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae319.

[11] Thomas E. Creighton. *Proteins: Structures and Molecular Properties*. W. H. Freeman, New York, 2 edition, 1993.

[12] Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer. Preprint.

[13] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D. Boscaini, M. M. Bronstein, and B. E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[14] Ketan S. Gajiwala, Jin Feng, Raul Ferre, Kristen Ryan, Olga Brodsky, Scott Weinrich, John C. Kath, and Andrew Stewart. Insights into the aberrant activity of mutant egfr kinase domain and drug recognition. *Structure*, 21(2):209–219, 2013. doi: 10.1016/j.str.2012.11.014.

[15] Bowen Gao, Yinjun Jia, Yuanle Mo, Yuyan Ni, Weiying Ma, Zhiming Ma, and Yanyan Lan. Profsa: Self-supervised pocket pretraining via protein fragment-surroundings alignment. *arXiv*, 2024. doi: 10.48550/arXiv.2310.07229.

[16] Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2024. doi: 10.1021/acs.jcim.3c01250.

[17] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: A deep learning library for drug-target interaction prediction. *Bioinformatics*, 2020.

[18] Gareth Jones, Peter Willett, Robert C. Glen, Andrew R. Leach, and Robert Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267: 727–748, 1997. doi: 10.1006/jmbi.1996.0897.

[19] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

[20] H. Lai, L. Wang, R. Qian, and et al. Interformer: an interaction-aware model for protein-ligand docking and affinity prediction. *Nature Communications*, 15(10223), 2024. doi: 10.1038/s41467-024-54440-6.

[21] Eunjoo Lee, Jiho Yoo, Huisun Lee, and Seunghoon Hong. MetaDTA: Meta-learning-based drug-target binding affinity prediction. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[22] F Leidner, N Kurt Yilmaz, and C A Schiffer. Target-specific prediction of ligand affinity with structure-based interaction fingerprints. *Journal of Chemical Information and Modeling*, 59(9): 3679–3691, 2019. doi: 10.1021/acs.jcim.9b00457.

[23] Krinos Li, Xianglu Xiao, Zijun Zhong, and Guang Yang. Accurate and generalizable protein-ligand binding affinity prediction with geometric deep learning. *arXiv*, 2025. doi: 10.48550/arXiv.2504.16261.

[24] Jiaqi Liao, Haoyang Chen, Lesong Wei, and Leyi Wei. Gsaml-dta: An interpretable drug-target binding affinity prediction model based on graph neural networks with self-attention mechanism and mutual information. *Computers in Biology and Medicine*, 150:106145, 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.106145.

[25] Pierre-Yves Libouban, Camille Parisel, Maxime Song, Samia Aci-Sèche, Jose C. Gómez-Tamayo, Gary Tresadern, and Pascal Bonnet. Spatio-temporal learning from molecular dynamics simulations for protein–ligand binding affinity prediction. *Bioinformatics*, 41(8):btaf429, 2025. doi: 10.1093/bioinformatics/btaf429.

[26] Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988, 2019.

[27] Tiqing Liu, Linda Hwang, Stephen K. Burley, Carmen I. Nitsche, Christopher Southan, W. Patrick Walters, and Michael K. Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(D1):D1633–D1644, 2024. doi: 10.1093/nar/gkae1075.

[28] Matthew R. Masters, Amr H. Mahmoud, and Markus A. Lill. Do deep learning models for co-folding learn the physics of protein-ligand interactions? *bioRxiv*, 2024. doi: 10.1101/2024.06.03.597219.

[29] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 10 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa921.

[30] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(33), 2011. doi: 10.1186/1758-2946-3-33.

[31] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

[32] G. A. Petsko and D. Ringe. *Protein Structure and Function*. New Science Press, 2004.

[33] J. Qiao, W. Gao, J. Jin, and et al. Molecular pretraining models towards molecular property prediction. *Sci. China Inf. Sci.*, 68:170104, 2025. doi: 10.1007/s11432-024-4457-2.

[34] Zachary A Rollins, Alan C Cheng, and Essam Metwally. Molprop: Molecular property prediction with multimodal language and graph fusion. *Journal of Cheminformatics*, 16(1):56, 2024.

[35] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *arXiv*, 2021. doi: 10.48550/arXiv.2106.09553.

[36] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

[37] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. *Proceedings of the 38th International Conference on Machine Learning*, 139:9323–9332, 2021. doi: 10.5555/3546022.3546193.

[38] Robert P. Sheridan, Prakash Karnachi, Matthew Tudor, Yang Xu, An Liaw, Fardok Shah, Alan C. Cheng, Emily Joshi, Merrill Glick, and Jose Alvarez. Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure-activity relationship models? *Journal of Chemical Information and Modeling*, 60(4):1969–1982, 2020.

[39] Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, et al. Chemberta-3: An open source training framework for chemical foundation models. *ChemRxiv*, 2025.

[40] Jeongtae Son and Dongsup Kim. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PLOS ONE*, 16(4):1–13, 04 2021. doi: 10.1371/journal.pone.0249404.

[41] Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pages 20479–20502. PMLR, 2022.

[42] Claire Suen and Alan Cheng. Integrating bilinear transduction with message passing neural networks for improved admet property prediction. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-dm67f. Preprint.

[43] Anqi Wang, C. David Stout, Qi Zhang, and Eric F. Johnson. Contributions of ionic interactions and protein dynamics to cytochrome p450 2d6 (cyp2d6) substrate and inhibitor binding. *The Journal of Biological Chemistry*, 290(8):5092–5104, 2015. doi: 10.1074/jbc.M114.627661.

[44] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

[45] Zhiqi Xie, Peng Zhang, Zipeng Fan, Qingpeng Zhang, and Qianxi Lin. Em-pla: environment-aware heterogeneous graph-based multimodal protein–ligand binding affinity prediction. *Bioinformatics*, 41(7):btaf298, 2025. doi: 10.1093/bioinformatics/btaf298.

[46] Ziduo Yang, Weihe Zhong, Qiujie Lv, Tiejun Dong, and Calvin Yu-Chian Chen. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The Journal of Physical Chemistry Letters*, 14(8):2020–2033, 2023.

A Appendix

A.1 Additional methods

A.1.1 Datasets

The EGFR dataset was sourced from BindingDB (curated from literature, PubChem, patents/WIPO, and ChEMBL), containing $\sim 9,500$ molecules [27]. The Cytochrome P450 2D6 (CYP2D6 Inhibition) dataset was sourced from BindingDB (curated from literature, PubChem, patents/WIPO, and ChEMBL), containing $\sim 5,000$ molecules [19]. All train/val/test were obtainted through a random split.

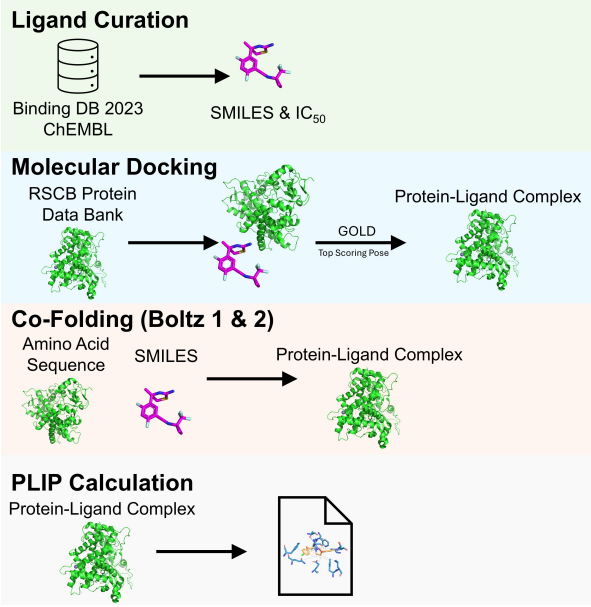


Figure A1: Overview of Molecular Docking and Interaction Profiling Workflow.

A.1.2 Input preparation

For each ligand, its SMILES representation and corresponding IC50 value were obtained. Concurrently, the 3D atomic coordinates of the target proteins, EGFR and CYP2D6 Inhibition, were retrieved from the Protein Data Bank (PDB) [5, 14, 43]. Following retrieval, the protein structures were isolated, removing any co-crystallized ligands, water molecules, and unwanted residues. Both the isolated ligands and proteins were processed to add missing hydrogen atoms and further steps for subsequent docking simulations (Figure A1).

A.1.3 Additional details on Boltz structure prediction

The tool was installed from the original Git repository (version 2.2.0, commit c9b6af1). For each protein-ligand pair, a single diffusion sample was generated using a fixed seed of 42 and a step scale of 1.5, with all other parameters set to their default values. Specifically, the Boltz-1 model was run at precision 32, while the Boltz-2 model utilized mixed precision to optimize performance. A Multiple Sequence Alignment (MSA) was prepared using HHblits, and the ligand was provided to the models as a SMILES string. For the MOE post-processing steps, we used the Amber:EHT force field. To maintain the relative position of the ligand within the binding pocket, a ligand tether was applied. Additionally, atoms beyond 7Å from the ligand were fixed during the energy minimization to prevent unnecessary conformational changes in the peripheral regions of the protein.

A.1.4 Feature Selection

A.1.5 Ensembled predictions

For the docking-based models, predictions were obtained by averaging the outputs from models trained on multiple representative protein structures. For the CYP2D6 Inhibition dataset, we ensembled predictions from two separate models, each trained on a different RCSB protein structure. Due to computational constraints, a single RCSB structure was used for the EGFR dataset. For the Boltz and baseline Chemprop (D-MPNN) models, a two-fold, two-ensemble approach was used, resulting in four total model predictions per molecule. These four predictions were then averaged to produce a single final output. Standard error and standard deviation were calculated from the individual metrics of the four models.

A.2 Random Forest benchmarking methods for low-data regime

The Random Forest (RF) models were trained using the scikit-learn package. For feature representation, we used Morgan Fingerprints with a radius of 2 and 2048 bits. RF models were trained with 500 trees with the standard error and deviation calculated across each tree.

A.2.1 MoLFormer benchmarking methods for low-data regime

To benchmark against MoLFormer, we first extracted the pre-trained weights from the model. Using these weights, we performed hyperparameter optimization on our various training sets to find the best hyperparameters, as listed in Table A1. Next, we fine-tuned the model on these same training molecules. Then we evaluated the fine-tuned model's predictive performance on the test set. The hyperparmeters used for each training size are reported in Table A2.

Table A1: Hyperparameters Search Space for CYP2D6 Inhibition $n = 250, 500, 750$.

Hyperparameter	Search Space
Learning Rate (Head)	$\eta_h \in [10^{-6}, 10^{-3}]$
Learning Rate (Base)	$\eta_b \in [10^{-8}, 10^{-5}]$
Dropout	$p \in [0.2, 0.4]$
Number of Layers	$L \in \{2, 3, 4\}$
FFN Dim	$d_{ffn} \in \{64, 128, 256, 512\}$
Batch Size	$B \in \{4, 8, 16, 32\}$
Early Stopping	$P = 3$

Table A2: Optimized Hyperparmeters for MoLFormer finetuned on CYP2D6 Inhibition $n = 250, 500, 750$.

Hyperparameter	$n = 250$	$n = 500$	$n = 750$
Learning Rate (Head)	0.0006	0.0007	0.0007
Learning Rate (Base)	$5.67 * 10^{-7}$	$5.27 * 10^{-8}$	$7.57 * 10^{-7}$
Dropout	0.37	0.39	0.37
Number of Layers	2	3	3
FFN Dim	128	512	64
Batch Size	4	8	8
Epochs	50	50	50
Early Stopping	5	5	5

A.3 Additional figures and results

Table A3: Hyperparameter Information for D-MPNN models. We use the Chemprop hyperparameters described in our previous work [3, 8].

Hyperparameter	Value
MPN depth	4
MPN hidden size	600
FFN number of layers	4
FFN hidden size	1300
Dropout	0
Aggregation	Norm
Number of folds (training/validation split seed)	2
Ensemble size (parameter initialization seed)	2
Epochs	60

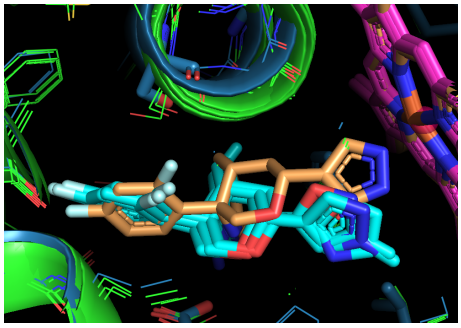


Figure A2: Structural validation of predicted poses. Structural Validation of Predicted Poses. We compared predicted binding poses for compounds similar to the known RCSB PDB structures. The captured images highlight that among the methods evaluated, the Boltz-2 model consistently produced the most accurate alignment to the PDB structures, particularly for the core substructure of each compound. This suggests that the improved predictive performance of the Boltz-2 model may be attributed, at least in part, to its ability to generate more physically realistic and well-aligned binding conformations.

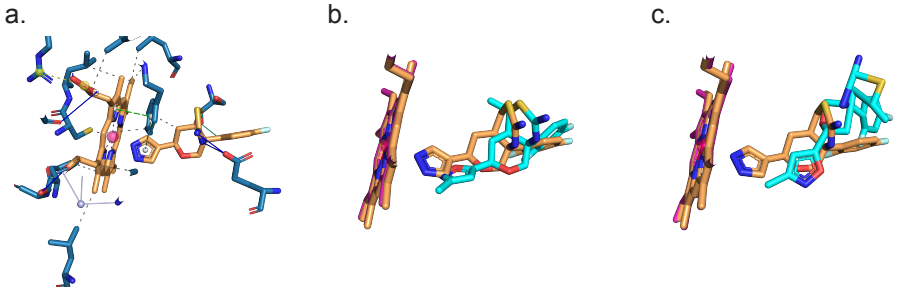


Figure A3: Structural validation of predicted poses for Compound 1790. Overlay predicted binding poses for compounds in training dataset similar to the known RCSB PDB structures (4XRZ) exhibiting π-stacking interactions, based on Tanimoto similarity. **a.** PLIF for 4XRZ **b.** Overlay of the Boltz-2 predicted structure (cyan) with 4XRZ (orange), identified π-stacking interactions **c.** Overlay of the Boltz-1 predicted structure (cyan) with 4XRZ (orange)

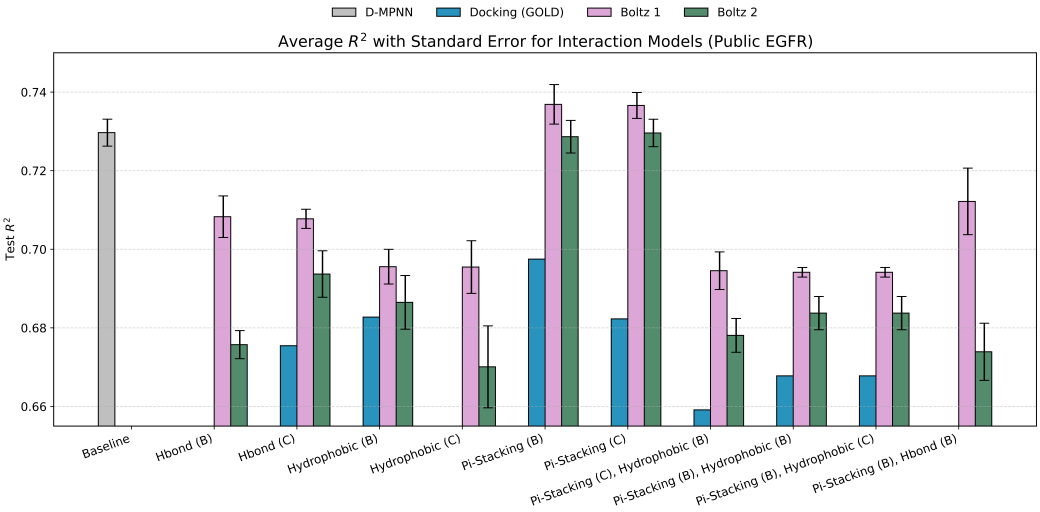


Figure A4: Performance of PLIP-Informed Models for Public EGFR. The bar chart compares the average R^2 with standard error for the baseline D-MPNN (Chemprop) model against selected PLIP-informed models.

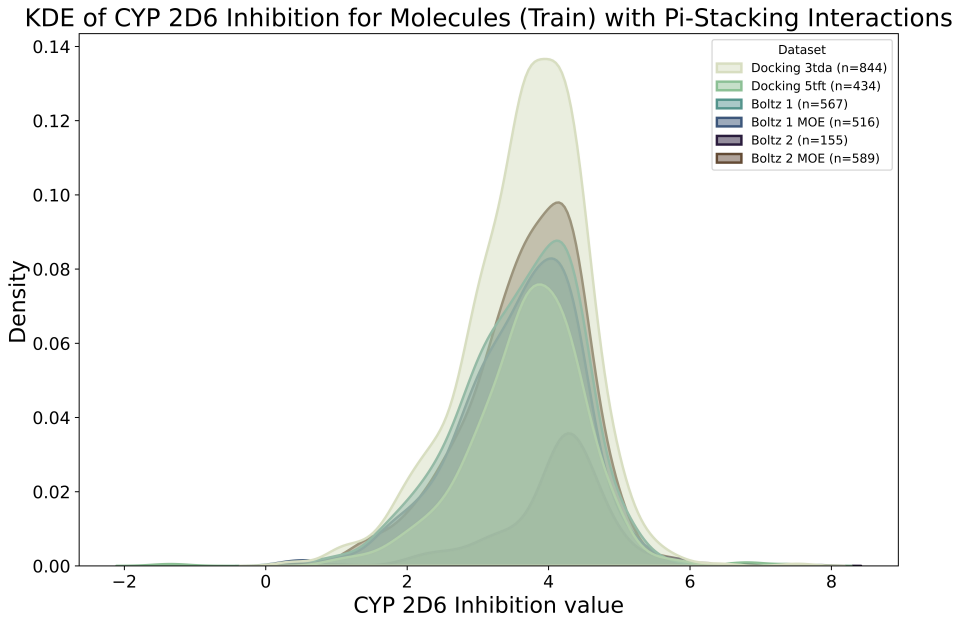


Figure A5: Distribution of π-Stacking Interactions in CYP2D6 Inhibition Train Set. Kernel Density Estimation (KDE) plot illustrating the distribution of $log(IC_{50})$, where IC_{50} values are in nM affinity units, that have an identified π-stacking interaction. The KDE plots for each pose-generation method (Boltz, Docking, etc) show that while the total number of identified π-stacking interaction varies significantly across methods, the overall distribution of these interactions mirror the left-tailed distribution of the full CYP2D6 dataset.

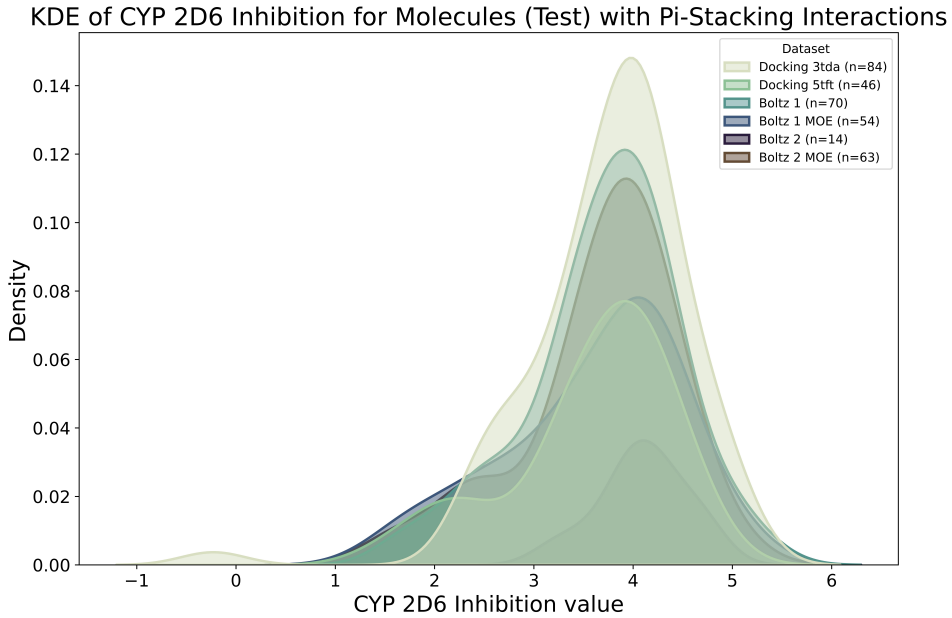
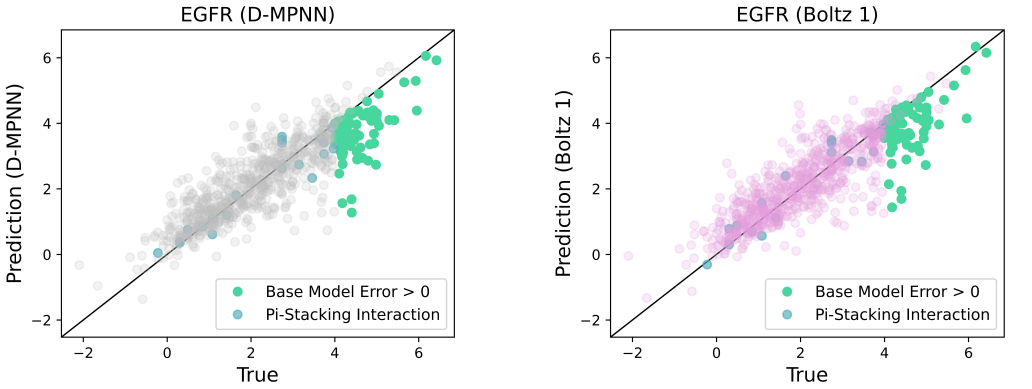


Figure A6: Distribution of π-Stacking Interactions in CYP2D6 Inhibition Test Set.



(a) DMPNN Parity Plot

(b) Boltz-1 (Continuous π-Stacking) Parity Plot

Figure A7: Comparison of Predicted versus True Potency for EGFR.

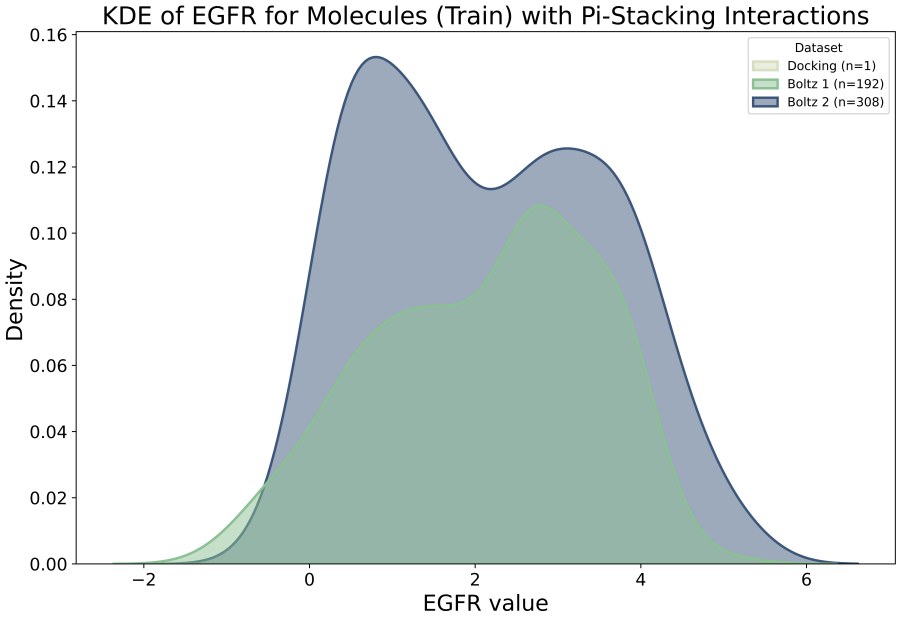


Figure A8: Distribution of π-Stacking Interactions in EGFR Train Set.

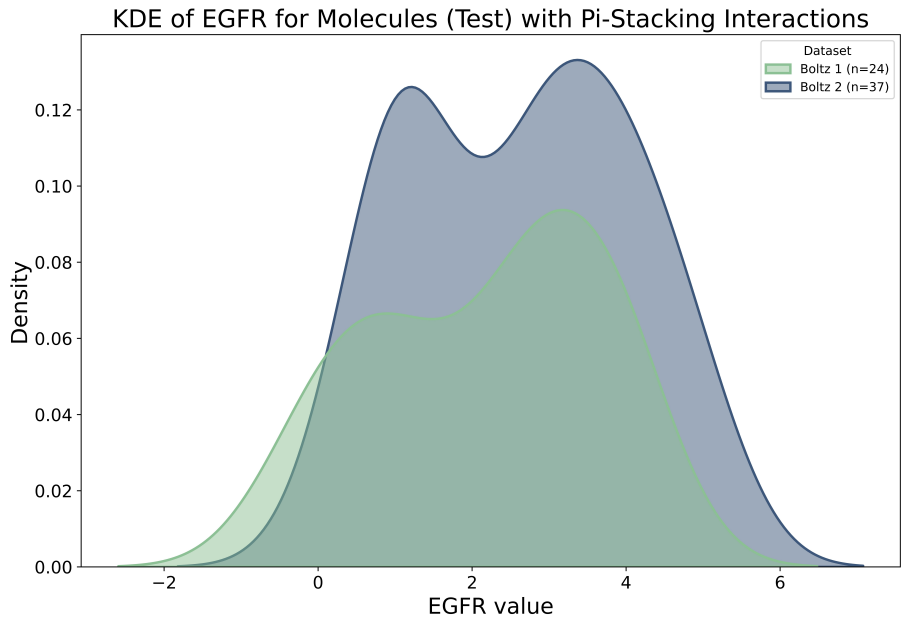


Figure A9: Distribution of π-Stacking Interactions in EGFR Test Set.

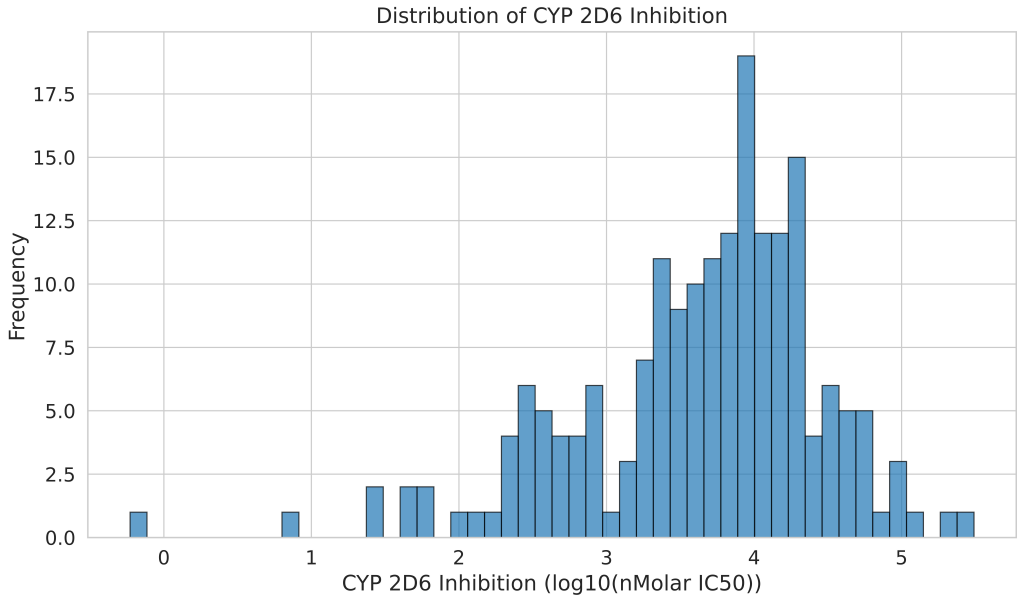


Figure A10: Distribution of CYP2D6 Inhibition test data.

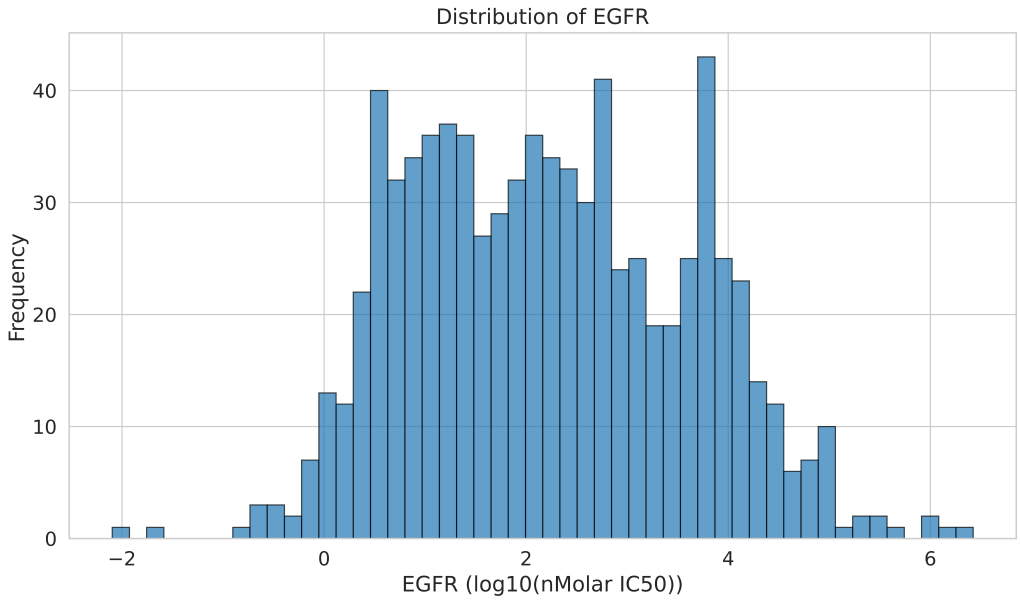


Figure A11: Distribution of EGFR test data.